Mathematical Challenges of the 21st Century

High-Dimensional Data Analysis: The Blessings and Curses of Dimensionality

David Donoho, Stanford
http://www-stat.stanford.edu/~donoho
Dr. Emmanuel Candès, CalTech
Dr. Xiaoming Huo, Georgia Tech
Dr. Arne Stoschek, Stanford
Ofer Levi, Stanford

Morning August 8, 1900

Figure 1: Hilbert in 1900

Morning August 8, 2000

Primary Risks

- Pretense
- Heaviness
- Incomprehensibility

Methodology for Today

- Broad Trends, Operating for decades
- Accelerating in Next Century
- Mathematical Opportunities

John Wilder Tukey, 1915-2000

Figure 2: Picture in Time/Life *Mathematics*

Tukey's Achievements

- English Language: 'Software', 'Bit'
- Computation: Fast Fourier Transform/Applications
- Mathematics: Ultrafilters, Axiom of Choice
- Statistics:
  - Robust Statistics
  - Multiple Comparisons
  - Spectrum Analysis
  - High Dimensional Data Analysis

Tukey's Achievements (ctd.)

- High Dimensional Data Analysis
  - Projection Pursuit
  - High-Dimensional Visualization
- Signal Processing
  - Cepstrum
  - Nonlinear Filtering (medians)
- Recognitions
  - Bell Labs, Princeton, Governmental Advisor
  - IEEE Medal, National Medal of Science

Tukey's Idiosyncracy

- Home Schooled
- Chemist First
- Compulsive Naming & Renaming
- Required translation

- Lionel Trilling's Definition of Genius
  - Unique Personality
  - Expresses Personality through intellectual output
  - Neither right nor wrong
- Contrast to Mathematician's notion of Genius

Tukey's Apogee – Early 1960's

- Fast Fourier Transform
- Robust Regression, Ubiquity of NonGaussianity
- Nonlinear signal processing
- Data Analysis vs. Math. Stat.

Data Analysis

- Future of Data Analysis 1962
  - Data are coming
  - Must analyze data even in homely ways
  - Math can get in the way
- Data Analysis Including Statistics 1968
  - Data analysis is huge activity
  - Growth towards massive data
  - Statistics tiny subset

- Exploratory Data Analysis 1976
  - Deal directly with data
  - Even by hand ...
- Not particularly welcome
  - Controversy in 1962
  - Reception in 1977

Passages from Future of Data Analysis

“We dare not neglect any of the tools that have proved useful in the past. But equally we dare not find ourselves confined to their use. If algebra and analysis cannot help us, we must press on just the same, making as good use of intuition and originality as we know how.” ”We need to face up to more realistic problems.”

“...data analysis is intrinsically an empirical science.”

“Dare we adventure?”

“Who is for the challenge?”

What has Happened

- Data Analysis: separate field
  - 100,000’s of practitioners
  - Biology: Genomics
  - Finance: Program Trading
  - Marketing: Data Mining
- Engine: Information Technology
  - Moore’s Law
  - Disk Drives
  - Internet
  - sensors

- Mathematics: relatively small interaction
  - Scandalized by Tukey’s initiative
  - Perception: little intellectual content

New York Times Obituary

Emphasis not on Tukey’s mathematical achievements. Instead: a respected mathematician **turning his back on proof**, focusing **on analyzing data** rather than **proving theorems**.

> “He legitimized that, because he wasn’t doing it because he wasn’t good at math,” Mr. Wainer said. “He was doing it because it was the right thing to do.”

NYT July 28, 2000

My Thesis

Tukey was right in 1962 & some points still right

- Data Analysis is activity of immense importance
- Data Analysis will continue growing dramatically
- Data Analysis poses major challenges
- Computing advances major engine of progress

Separation from mathematics **no longer makes sense**

- Computing advances have run their course
- Fundamental Roadblocks **only** Mathematical
- Payoffs large and widespread (Society & Math)

Important to Recall

Most fundamental data analysis tools in use today based on work before schism

- Regression
- Classification
- PCA

Historically, talented developers had deep mathematical connections

- Gauss, Laplace, ...
- Fisher

- Hotelling

Particularly, insights into high-dimensional space...

Volume of Tubes story

(a) Hotelling

(b) Weyl

Agenda For Today

- Data
- Data Structures
- Data Analysis
- Centrality of High-Dimensions
- Curse of Dimensionality
- Blessings of Dimensionality
- Mathematical Challenges

I. Data

- Sensor
- Financial
- Imagery
- Hyperspectral
- Gene Expression
- Data Availability
- Mirror Worlds

Sensor Data – EEG Array

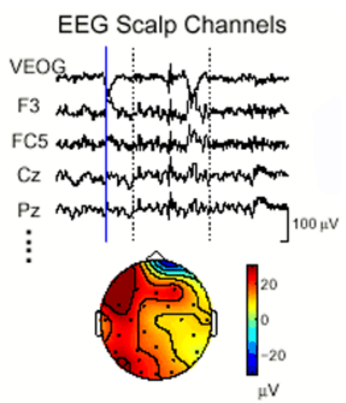


Figure 3: EEG Array Data; Jung, Sejnowski

Image Data – Facial Gestures

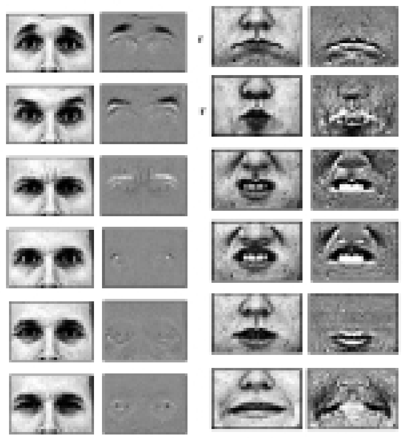
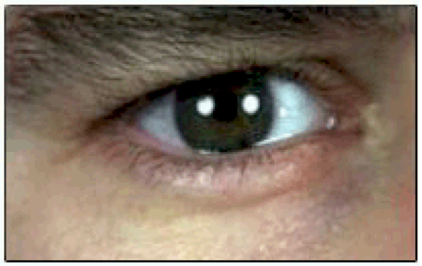
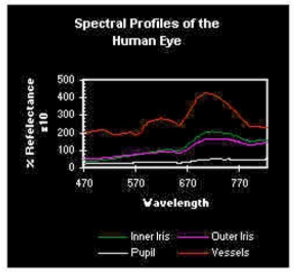


Figure 4: Facial Gesture Data

Hyperspectral Photography



(a) Eye



(b) Spectrum

Hyperspectral Photography – Derived Image

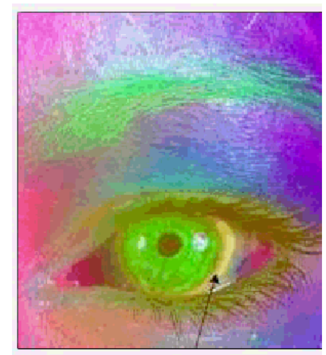
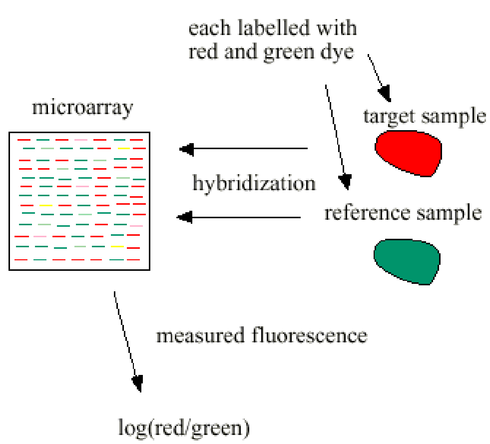


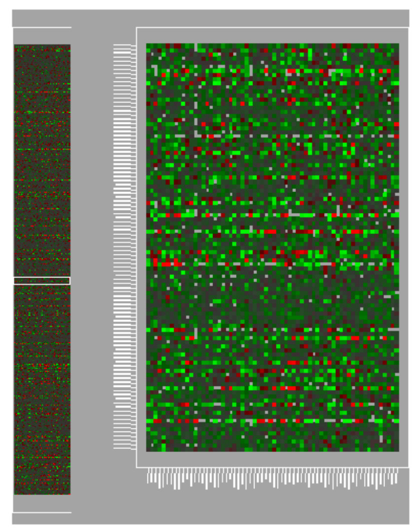
Figure 6: Derived Image

Gene Expression Experiment



Hastie, Tibshirani, et al. (2000)

Gene Expression Data



Important Trends

- Data availability
- Mirror Worlds

Financial Data Availability

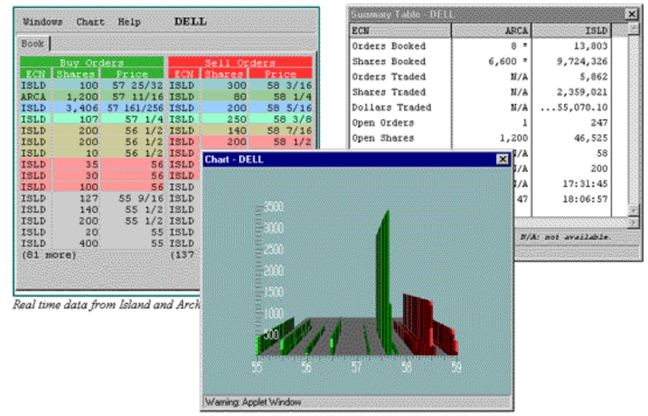
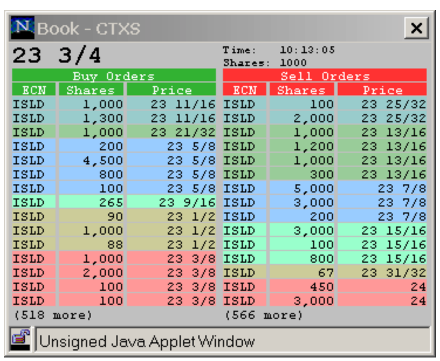


Figure 7: www.Island.com

Financial Data – Bids



Book - CTXS

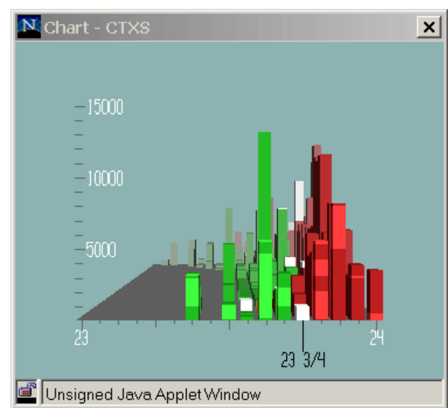
23 3/4

Time: 10:13:05
Shares: 1000

Buy Orders			Sell Orders		
ECN	Shares	Price	ECN	Shares	Price
ISLD	1,000	23 11/16	ISLD	100	23 25/32
ISLD	1,300	23 11/16	ISLD	2,000	23 25/32
ISLD	1,000	23 21/32	ISLD	1,000	23 13/16
ISLD	200	23 5/8	ISLD	1,200	23 13/16
ISLD	4,500	23 5/8	ISLD	1,000	23 13/16
ISLD	800	23 5/8	ISLD	300	23 13/16
ISLD	100	23 5/8	ISLD	5,000	23 7/8
ISLD	265	23 9/16	ISLD	3,000	23 7/8
ISLD	90	23 1/2	ISLD	200	23 7/8
ISLD	1,000	23 1/2	ISLD	3,000	23 15/16
ISLD	88	23 1/2	ISLD	100	23 15/16
ISLD	1,000	23 3/8	ISLD	800	23 15/16
ISLD	2,000	23 3/8	ISLD	67	23 31/32
ISLD	100	23 3/8	ISLD	450	24
ISLD	100	23 3/8	ISLD	3,000	24
(518 more)			(566 more)		

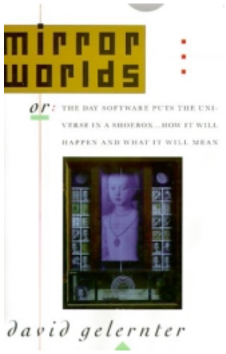
Unsigned Java Applet Window

Financial Data – Charts



Mirror Worlds

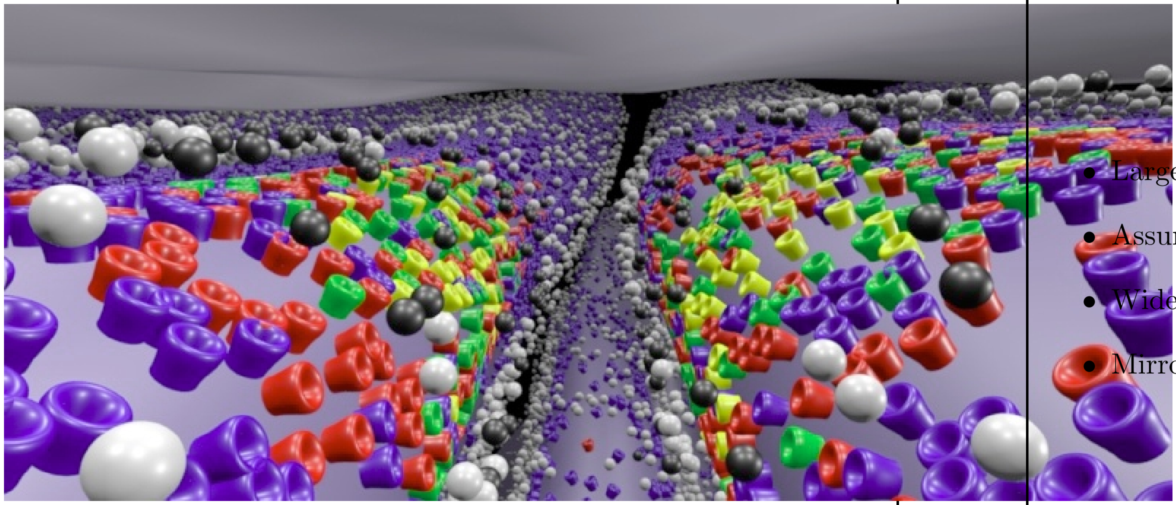
(a) Gelertner



(b) Book

www.e-cell.org

www.mcell.cnl.salk.edu



Conclusions about Data

- Large Scale Inundation
- Assumed useful
- Widespread availability
- Mirror Worlds

II. Data Structures

Many important data sources can be captured in the format of N by D array:

$$X_{i,j} : \quad 1 \leq i \leq N, \; 1 \leq j \leq D$$

Here

- Row: observation of all attributes on one subject
- Col: observation of all subjects on one attribute

Example: Term-Document Matrices

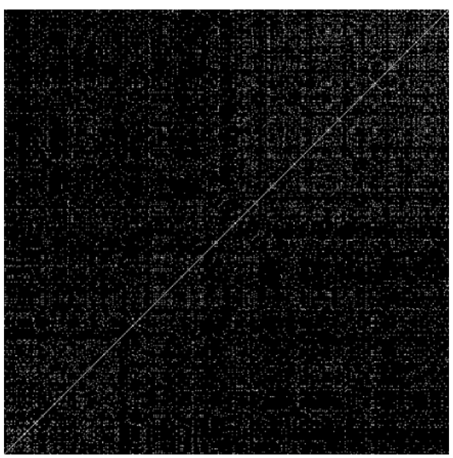
Example: Murtagh, Starck, Berry (2000)

- Journal *Astronomy and Astrophysics*
- $N = 512$ articles
- $D = 269$ terms
- $X_{i,j}$ frequency of terms in articles (normalized)

Here

- Row: rel frequencies of all terms on one articles
- Col: rel frequencies of one term on all articles

Term Document Matrix Example



III. Data Analysis

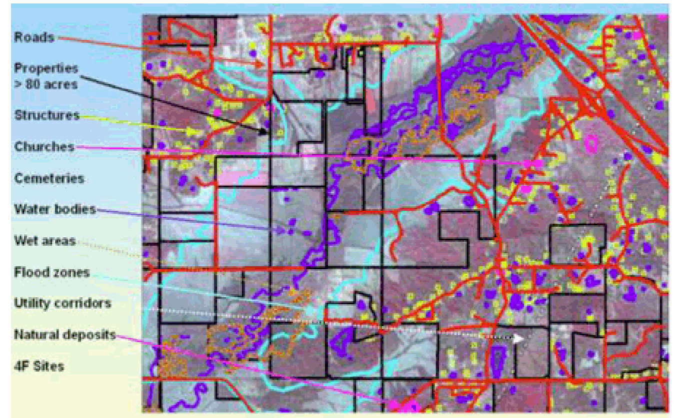
Many standard data analysis tasks formulated in terms of N by D array. Here

- Classification
- Regression
- Hidden Components Analysis
- Clustering

Classification

- Categorical variable: $X_{i,1}$
- Predictor variables: $X_{i,j}$, $j > 1$.
- Predict category from other variables
- Fisher Linear Discrimination
- k-nearest neighbor

Classifier Output on Hyperspectral Imagery



Regression

- Dependent variable: $Y_i \equiv X_{i,1}$; Predictor variables: $X_{i,j}$, $j > 1$.
- Predict Y from other variables X
- Linear regression

$$Y_i = a_1 + a_2 X_{i,2} + \ldots + a_D X_{i,D} + Z_i$$

- Methods: least squares, least absolute, robust ...
- Nonlinear regression

$$Y_i = f(X_{i,2}, \ldots, X_{i,D}) + Z_i$$

- Methods: nearest neighbors, neural nets, radial basis functions, projection pursuit ...

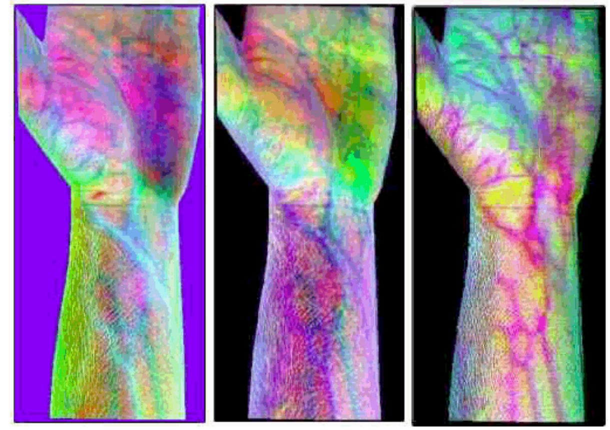
Hidden Components

- Model $X = AS$
- X – observed data
- A – linear transformation (unknown)
- S – latent or hidden components (unknown)
- Principal Components Analysis
- Factor Analysis
- Independent Components Analysis
- ...

Principal Components Analysis

- Empirical Covariance Matrix: $C = N^{-1} X^T X$
- Empirical Eigenbasis $U = [U_1 U_2 \dots U_D]$
- Empirical Eigenvalues $\Lambda = Diag(l_1, \dots, l_D)$.
- Factorization: $C = U \Lambda U^T$, U orthogonal
- Other Names: SVD, Karhunen Loève
- Factorization: $X = U \Sigma V^T$, U,V orthogonal, $\Sigma = \Lambda^{1/2}$ diagonal.

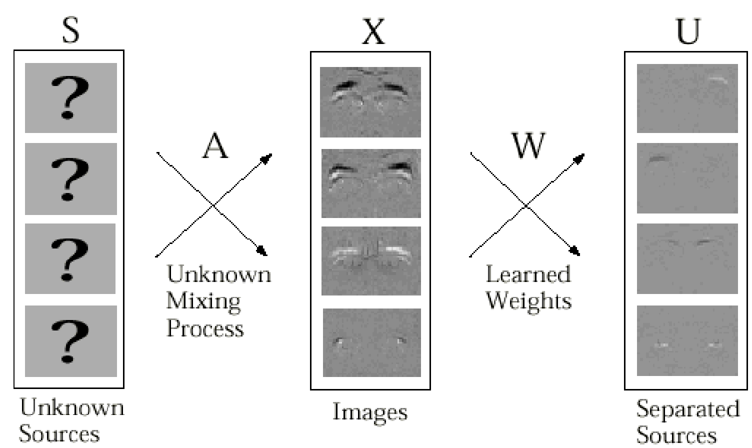
PCA Output on Hyperspectral Imagery



Independent Components Analysis

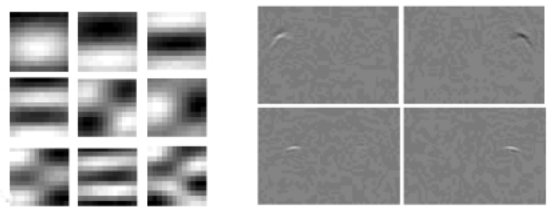
- Seeks factorization $X = AS$
- A not nec. orthogonal
- S sparse, nongaussian, independent
- Constellation of heuristic methods
  - Diagonalizing High-Order Cumulants
  - Maximum Nongaussianity
  - Minimum Mutual Information

ICA Model on Facial Gesture Data



Donato et al.(1999) (Sejnowksi lab)

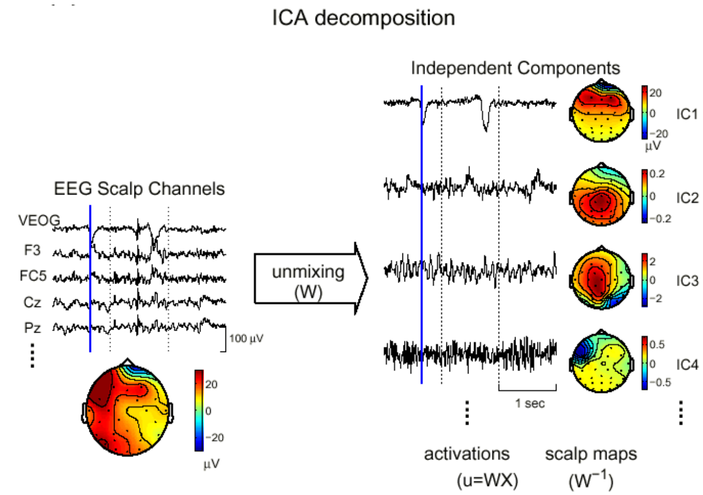
Comparison of PCA/ICA on Facial Gesture



(a) PCA (b) ICA

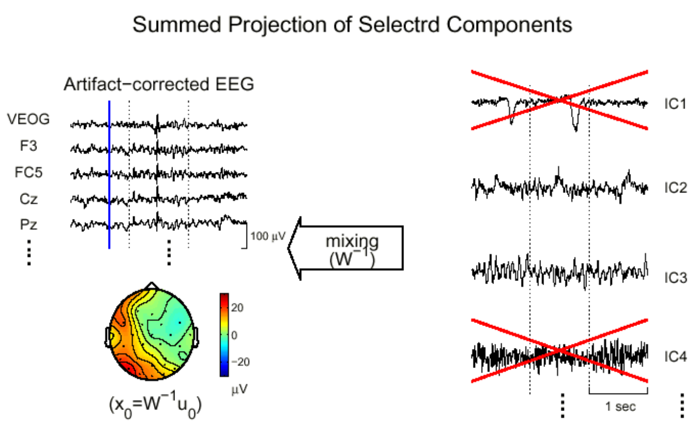
Donato et al.(1999) (Sejnowksi lab)

ICA Output on EEG Array Data



Jung et al. (2000) (Sejnowksi lab)

ICA Corrected EEG Array Data



Jung et al. (2000) (Sejnowksi lab)

Clustering

- Model rows, columns of X unordered
- Arrange so that nearby rows (columns) similar
- Large numbers of heuristic procedures

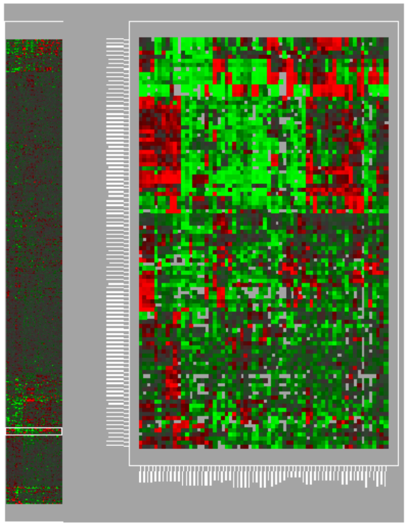
Gene Expression Data

- Order rows (genes) columns (cell lines)
- Tree methods
- Gene Shaving (Hastie, Tibshirani, et al.)
- Plaid Models (Lazzeroni and Owen)

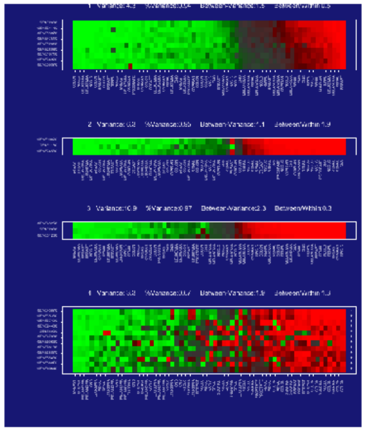
$$X = \mu_0 + \sum_{k=1}^{K} \mu_k \alpha_k \beta_k^T$$

- K clusters
- μ_k activation
- α_k binary N-vector

Gene Clustering Output



Gene Shaving Output



Hastie, Tibshirani, et al.

IV. High Dimensionality

- Think of N by D array as N points in D-space
- Tendency to large D
- Many measured attributes
- Automatically collected
- Presumably relatively few key attributes
- Unwilling to specify in advance

The Modern Difference

- Classically: D fixed, $N \to \infty$
- Modern Situation: $D = \beta N$
- Sometimes N fixed, $D \to \infty$
- Classically: all variables pertinent
- Modern Situation: unknown which pertinent
- Unwilling to specify in advance
- "Data Mining"

The Opportunity

- Need theory to surmount problems
- Theory for D very large
- Theory for few relevant variables
- Cope with Curses of Dimensionality
- Exploit Blessings
- Theory will be based on asymptotics & approximations, not exact geometry as earlier

V. Curse of Dimensionality

Figure 15: Richard Bellman

Areas where seen

- Optimization by Exhaustive search
- Integration over Product domains
- Approximation over high-dimensional domains

Basic point:

- Approximate optimization
- Function of D variables
- Domain $[0,1]^D$
- f Lipschitz
- Order $(1/\epsilon)^d$ evaluations to obtain error ϵ

Implication in Statistics:

- Statistical estimation
- Function of D variables
- f Lipschitz
- Need $N \asymp (1/\epsilon)^{(D+2)}$ observations for RMS error ϵ
- Proof: minimax decision theory, hypercubes.

VI. Blessings of Dimensionality

1. Concentration of measure
2. Asymptotic Distribution
3. Approach to Continuum

Blessing 1: Concentration of Measure

- P Uniform measure on $D-1$-dimensional sphere
- X distributed P
- $f()$ Lipschitz function on S^{D-1}

$$P\{|f(X) - Ef(X)| > t\} \leq C_1 \exp(-C_2 t^2). \tag{1}$$

- Very rapid decay of tails – similar to Gaussian

Variants of Concentration of Measure

Example

- Example, $X_1, \ldots, X_D$ i.i.d. $N(0,1)$
- $f()$ Lipschitz function on $\mathbf{R^D}$

$$P\{|f(X) - Ef(X)| > t\} \leq C_1 \exp(-C_2 t^2). \tag{2}$$

- Example: $f(X_1, \ldots, X_D) = \max(X_1, \ldots, X_D)$.
- $M_D = \max(X_1, \ldots, X_D)$.
- $EM_D < \sqrt{2\log(D)}$
- Conclude: $P\{M_D > \sqrt{2\log(D)} + t\}$ decays rapidly in t.

Blessing 2: Asymptotic Distribution

- T_D statistic on D-dimensional data
- $P(T_D < t)$ too complex to work with
- Limiting Distribution

$$P(a_D(T_D - b_D) \leq t) \rightarrow G(t)$$

- a_D, b_D scaling and centering constants.
- Examples
  - Central Limit Theorem
  - Extreme Value Distributions
- Typically in settings where variables exchangeable

Blessing 3: Approach to Continuum

- T_D statistic on D-dimensional data
- Variables samples of continuous stochastic proces

$$X_j = X(j/D)$$

- Continuous Stochastic process well-understood
  - Brownian Bridge
  - Brownian Motion
  - Stationary process
- Limiting Properties of discrete data derive from continuous stochastic process.

Approach to Continuum (ctd.)

- Brownian Bridge on $[0,1]$
- Eigenanalysis of Covariance: Sinusoids
- Approximate eigenvectors of empirical covariance: also sinusoids.

Blessings in Action

1. Model Selection
2. Top Eigenvalue
3. Diagonalizing Cumulant Form

Blessing 1 in Action

- High-Dimensional Regression

$$Y_i = a_1 + a_2 X_{i,2} + \ldots + a_D X_{i,D} + Z_i$$

- Assumed Sparse Effects
- Assumed $k \ll D$ explanatory variables relevant
- **Data Mining**

Penalized Model Fitting

$$\min \mathrm{RSS}(\mathrm{Model}) + \lambda \mathrm{Model\ Complexity},$$

- RSS sum of squares of residuals $Y_i - Model_{i,1}$,
- Model complexity # variables $X_{i,2}, \ldots, X_{i,D}$ in model.
- λ penalty factor
- Impose a cost on large complex models.
  - 1960's 1970's: $\lambda = 2 \cdot \sigma^2$
  - σ^2 assumed variance of noise
  - Today , $\lambda = 2 \cdot \sigma^2 \cdot \log(D)$

Logarithmic penalty Interpretation

- Adjustment to RSS if search D variables
- Charge $2\log(D)\sigma^2$ to RSS rather than $2\sigma^2$ to include a variable
- Rationale:
  - Oracle knows which variables important
  - How well can Data Miner do?
  - Theorem: Data Miner MSE always within factor $C \cdot \log(D)$ of ideal oracle MSE
- Surprise: $\log(D)$ vs. D^α (say)

Comments on Logarithmic Penalty

- Important: implies data mining **actually feasible**
- Sharp: optimality results indicate that this is the right penalty.
- Correct: optimality results indicate no other estimation strategy better.
- At base: **logarithmic penalty due to concentration of measure.**

References: Johnstone (1998), Birgé-Massart (1998)

Beyond ...

- Exploiting special structure of certain problems, penalty may be made smaller
  - False Discovery
  - Structured search (not all variables)
- Similar logarithmic penalties everywhere
  - Mixture of Experts
  - Portfolio Selection

Blessing 2 in Action

- High Dimensional Gaussian Data
- Is there a 1-dimensional subspace with high concentration?
- Again, “Data Mining”
- Behavior of top principal component $\lambda_1(C)$

Eigenvalues of Random Matrices

- Classical: Fixed D, Large N. T. Anderson (1950’s-1960’s)
- High-Dimensional Case: $D = \beta N$
  - Bulk Spectrum: Wigner semicircle Law (1950’s)
  - Top eigenvalue Gaussian Unitary Ensembles: Tracy & Widom (1990’s)
- Johnstone (2000): Top eigenvalue of empirical covariance matrix

Comments on Top Eigenvalue of Covariance

- Tracy-Widom: Painlevé Type II asymptotic distribution
- Johnstone: Empiricially useful already in dimension 6.
- Surprise: Easier to find results for top eigenvalue than bulk eigenvalues

Reference: Johnstone (2000)

Blessing 3: Approach to Continuum

- Goal: Independent Components Analysis
- Attempt to diagonalize 4-th cumulant tensor
- JADE: heuristic method Cardoso & Souloumiac (1993)
- Dataset: Ramp – artificial data with discontinuities.
- $N = D = 32$.
- Result: approximate diagonalizing bases

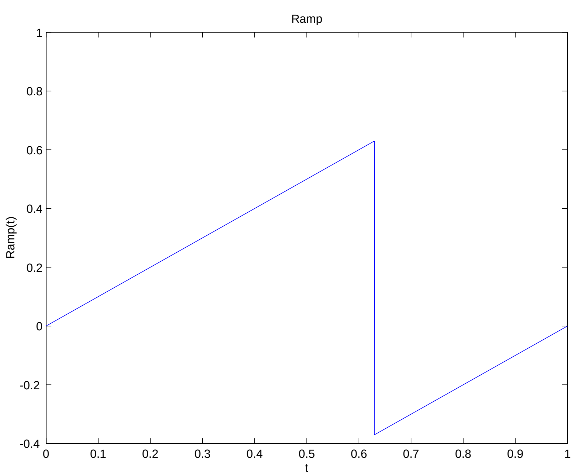
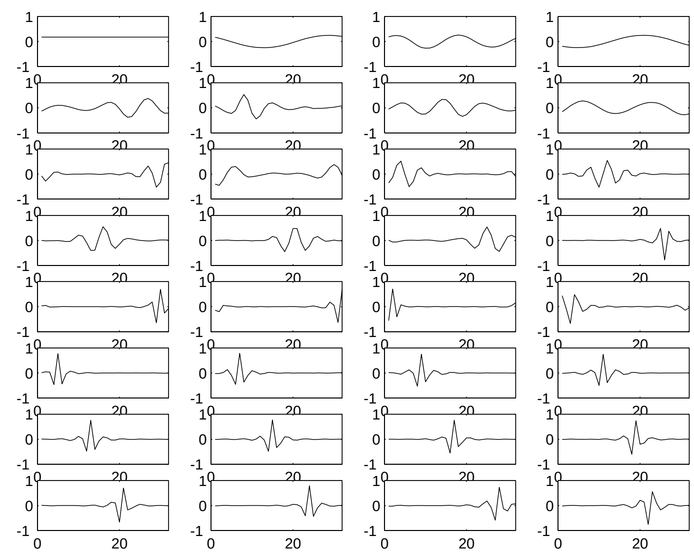
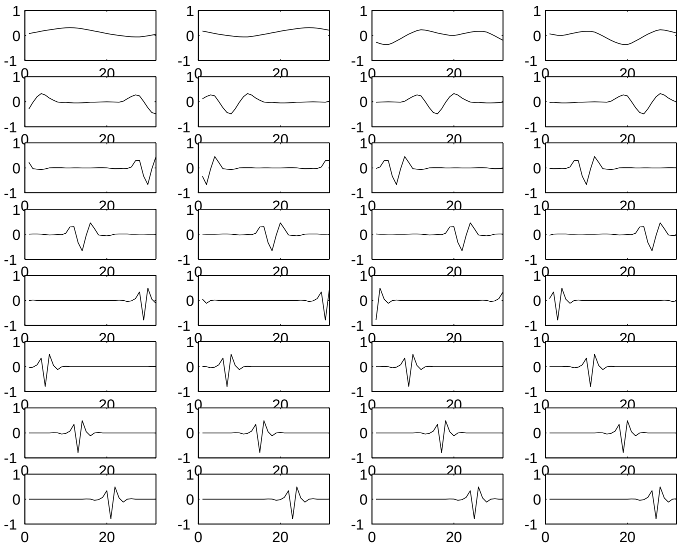


Figure 16: Typical Ramp

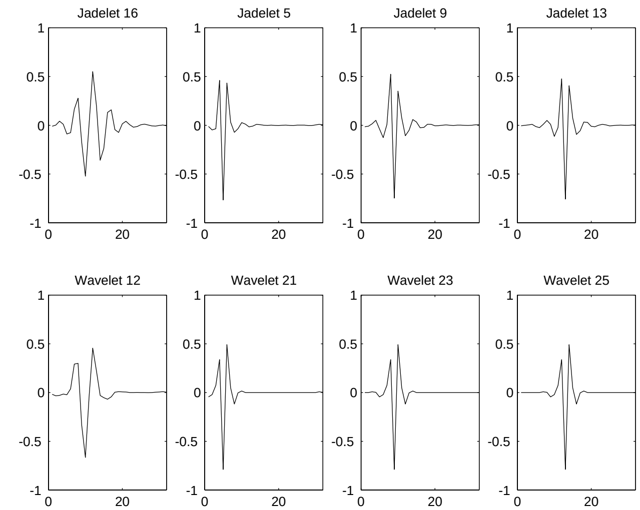
Results from JADE analysis of Ramp



Daubechies Nearly Symmetric Wavelets



Comparisons



Interpretations

Indep. Components discovered by JADE:

- Dyadic Scales
- Near-Translates
- Approx. Daubechies Nearly Symmetric

Approx. Solution: known basis from continuum theory

VII. Predictions:

Over the coming years:

- Attacks on Curse by refining assumptions about f
  - Example: Barron's results assuming $\nabla f \in \mathcal{F}L^1$.
  - Example: Coifman and J.O. Stromberg $\frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_D} f \in L^1$.
- Exploitations of Blessings
  - Example: Frieze & Kannan's fast approximate subspaces by randomized methods.
  - Example: Owen's fast test for approx. linearity

VIII. Personal Observations

Substantial room for synergy in coming years.

- High-dimensional data analysis can uncover existences of new mathematical objects (bases)
- Even Low-dimensional cases needs vigorous development both in math and in data analysis.

IX: Interpreting High-D Data Analysis

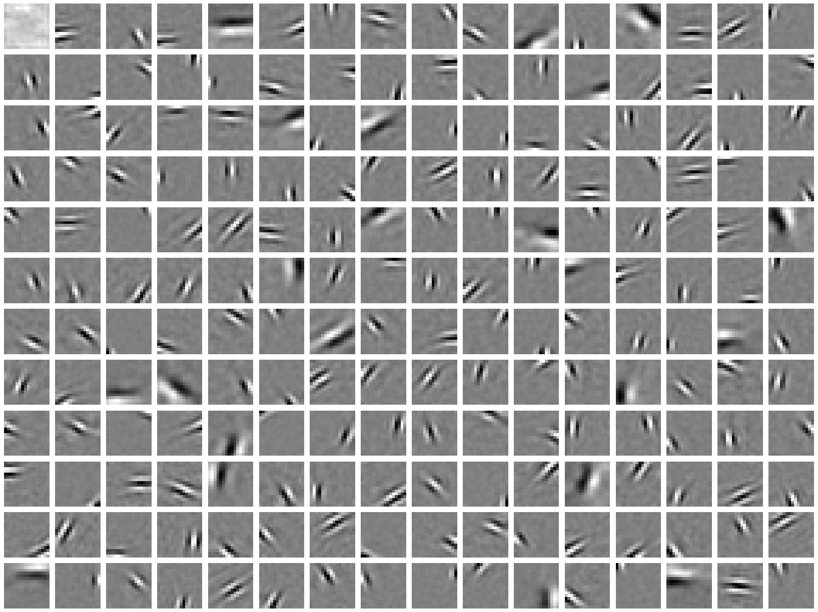
Vision Slogans

- 1970's: The eye does Fourier Analysis
- 1980's: The eye does Gabor Analysis
- 1990's: The eye does Wavelet Analysis
- Vision scientists use mathematics to provide language & framework.
- But what if mathematics did not yet create the right tools?

Challenging Output from High-D Data Analysis

Olshausen-Field Experiment

- Appeared in *Nature*, 1996
- Collection of natural images
- Extract 16×16 patches.
- Database of patches
- Indep. Components Analysis



Interpretation of Basis Elements

- Multi-Oriented
- Multiscale
- Bandpass

Initially, called basis elements “wavelets”. Later not in title.

My Reactions

- Not Fourier – too localized
- Not Gabor – too many scales
- Not Wavelets – too many directions
- What are these things?
- Mystery persists in $3-D$

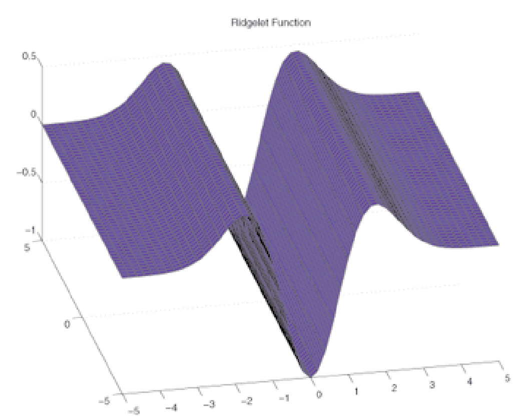
Some Recent Representations

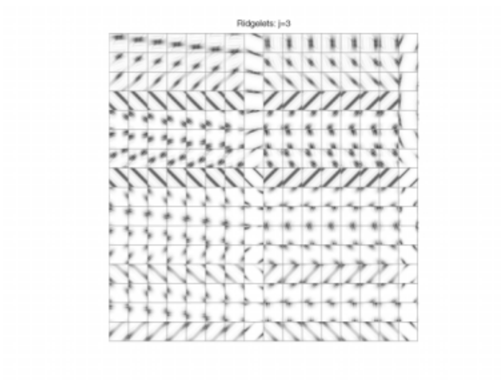
- Ridgelets – Candès Thesis (1998)

$$\rho(x; a, b, u) = \psi((u'x - b)/a)/a^{1/2}.$$

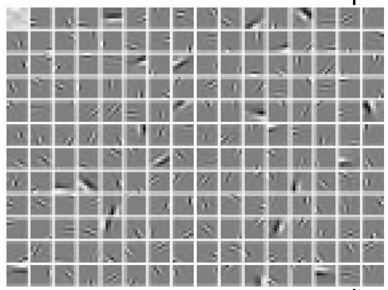
  - ψ 1-D wavelet
  - u unit vector, a scale, b location

- Curvelets – Candès & Donoho (1999)
  - multiscale ridgelets
  - $width = length^2$.
  - tight frame





(a) Ridgelets



(b) Imagelets

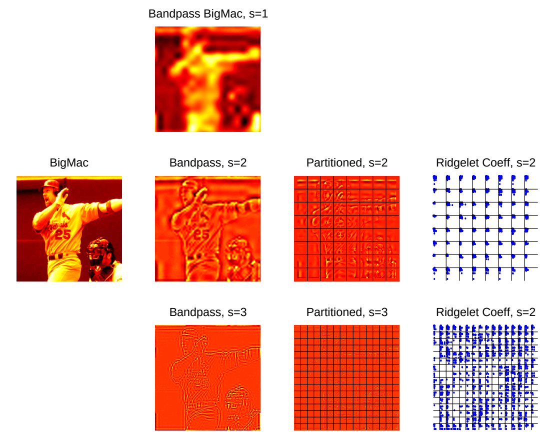


Figure 17: BigMac Image, and stages of Curvelet Analysis.

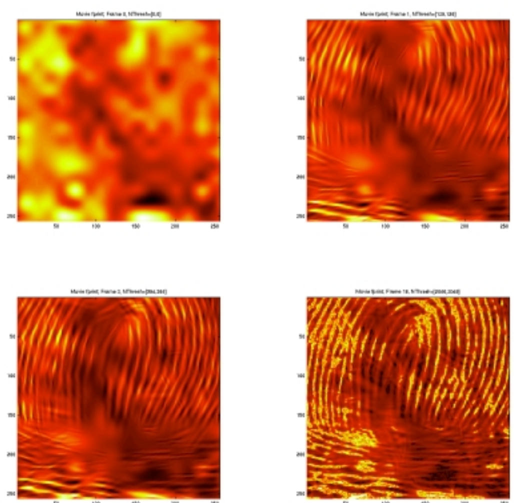


Figure 18: Fingerprint Movie. 64W + 0/256/768/3072 Curvelets

Figure 19: Lichtenstein, ‘In The Car’; 64 Wavelets+ 256 Curvelets

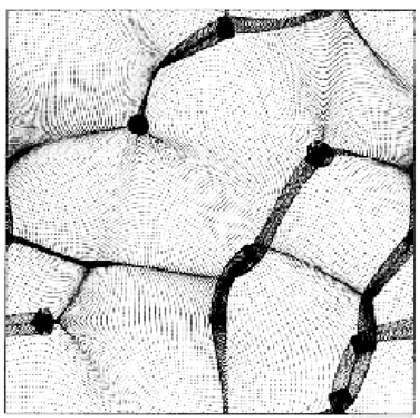
Comments

- Data analysis suggests existence of new harmonic analysis
- New harmonic analysis has surprising differences
- Data Analysis discovered **before** mathematicians

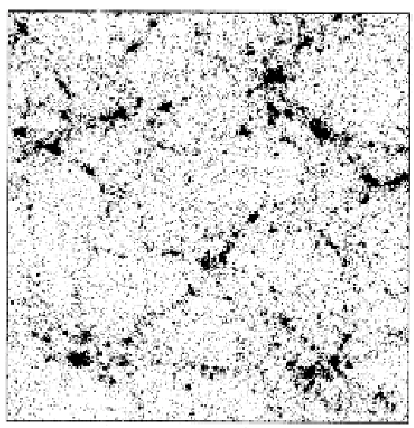
X: Frontiers in Low-Dimensional Data Analysis

- The Problem of Filaments
- The Travelling Salesman Problem
- Beamlets
- Applications

Filament Detection – Astronomy

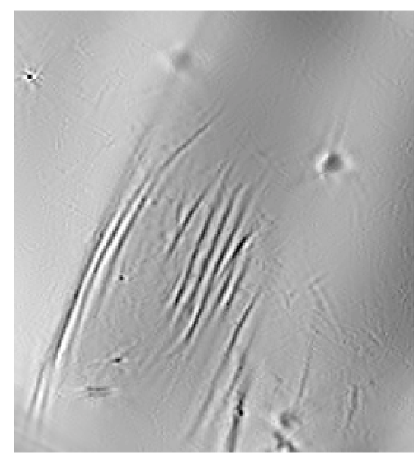


(a) Density



(b) Particles

Vesicle Detection – Electron Microscopy



Arne Stoschek et al.

Relevant Mathematics

Peter Jones Travelling Salesman Problem

- Countable set $S \subset [0,1]^2$
- Exists finite-length curve passing through points?

Solution by Geometric Analysis

- Q dyadic subsquare of $[0,1]^2$
- t_Q thickness of strip containing $S \cap 3Q$
- $\beta_Q = t_Q/\ell(Q)$, ℓ sidelength of Q
- **Theorem** *existence of a rectifiable curve containing S if and only if $\sum \beta(Q)^2 \ell(Q)^2$ finite.*

Aftermath

- Extensive generalization by G. David and S. Semmes
- Applications by Gilad Lerman to data analysis

My Reaction

- Part of a larger story?
- Systematic data structures?
- Systematic computational tools?

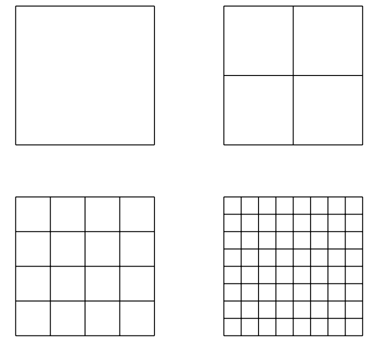
Recent Work (w/Xiaoming Huo)

- Beamlet Dyadic Pyramid
- Beamlet Graph
- Beamlet data structures
  - Network Flow Algorithms
  - Beamlet-decorated Partitioning

Beamlets (1)

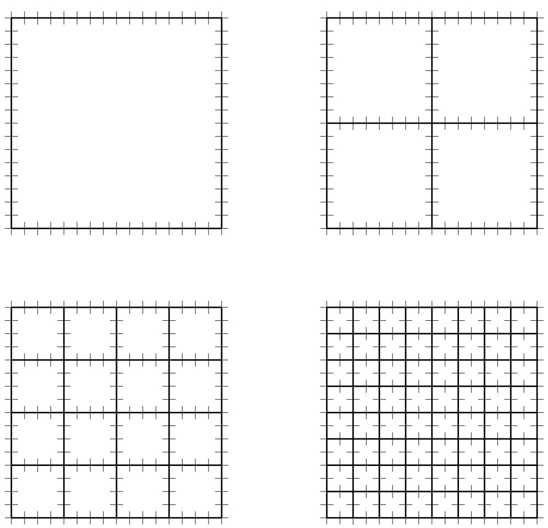
$\mathcal{B}_{n,\delta}$; $n = 2^J$, $\delta = 2^{-J-K}$, $K > 0$.

Start with Dyadic Squares in $[0,1]^2$



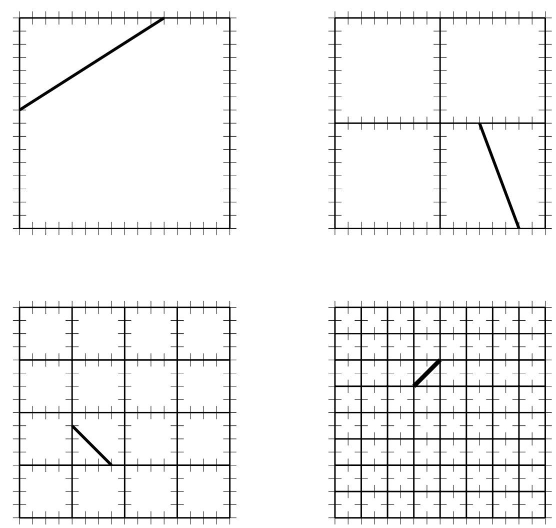
Beamlets (2)

Mark out all vertices with spacing δ

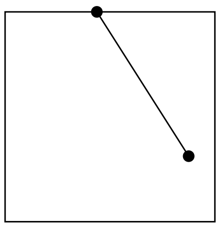


Beamlets (3)

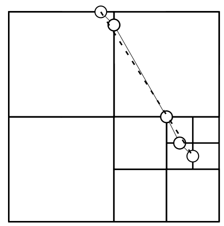
Connect all pairs of vertices in same square using line segments



Sparse Rep. of Line Segments by Beamlets



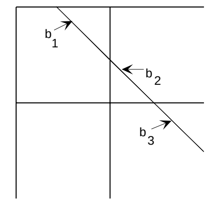
beam



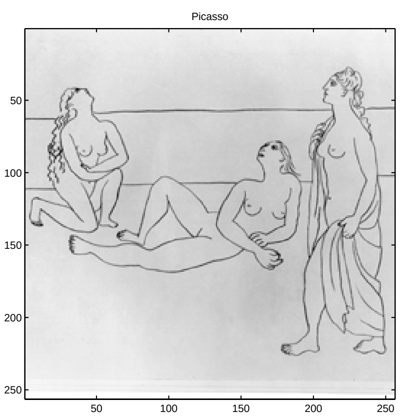
Chain of Beamlets

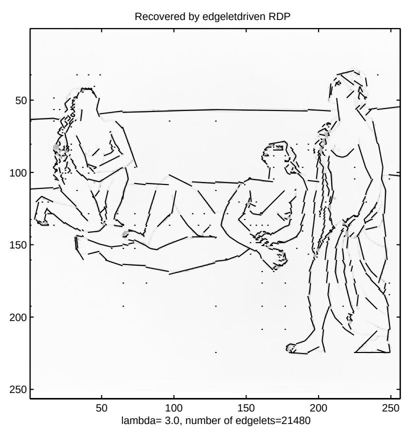
Two-Scale Relation

e decomposed as $\cup_i e_{,i}$ from finer level



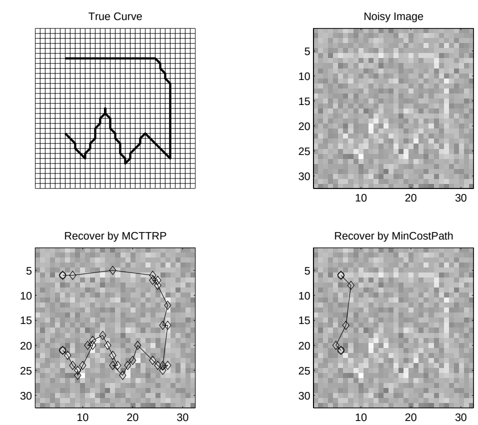
Extracting Lines





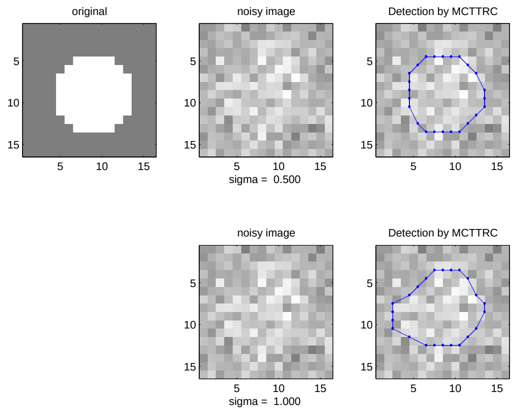
Xiaoming Huo

Filament Extraction by Network Flow



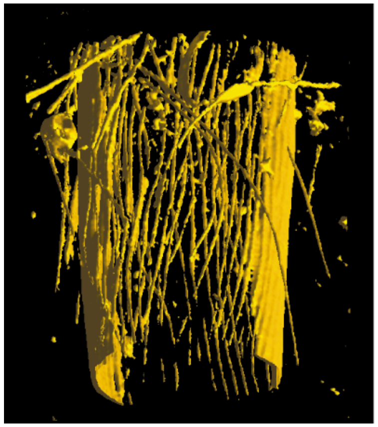
Xiaoming Huo

Object Extraction by Network Flow



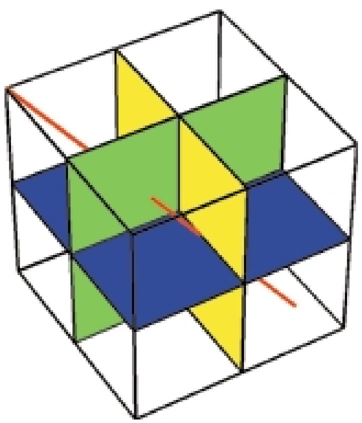
Xiaoming Huo

Importance of 3-D



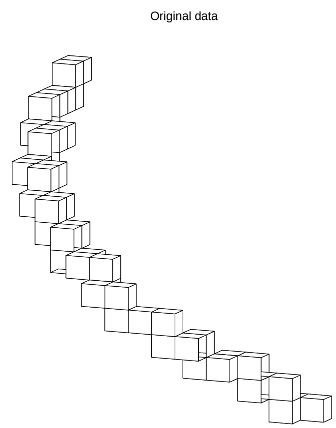
Arne Stoschek

Beamlets in 3-d

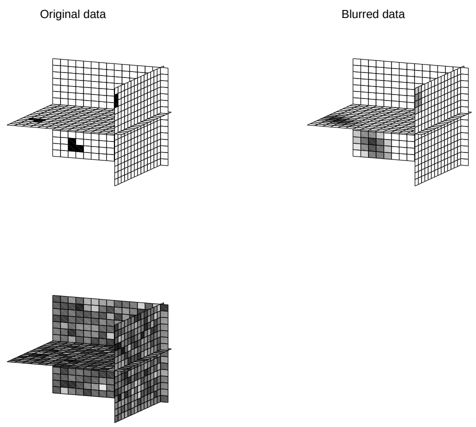


Ofer Levi

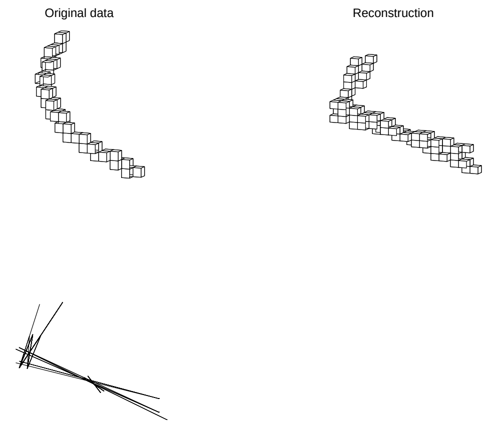
Object in 3-d



2-d Slices



Reconstruction



Summary on Low-Dimensional Geometry

- Need for local k-plane approx in $\mathbf{R}^{\mathbf{D}}$
- Mathematical need: study of Singular Integral Operators
- Data analysis need: filaments, curves in noisy data
- Synergistic development called for

Conclusion

- Data Analysis growing at breakneck speed
- Infrastructure of modern data analysis: classical
- Mathematical Tools for post classical infrastructure: instant, global Impact
- Data Analysis can uncover need for new Mathematical Analysis tools

Tukey, again

From a mathematical viewpoint, much of DLD youth was 'wasted' on data analysis

This has been redeemed.

DLD lived to see data analysis uncovering new mathematical structures – proto-ridgelets / proto-curvelets.

Thanks to John, and to the data analysts –

Olshausen and Field, Bell and Sejnowski van Hateren and Ruderman ...

Hilbert, again

... the question is urged upon us whether mathematics is doomed to the fate of those other sciences that have split up into separate branches, whose representatives scarcely understand one another and whose connection becomes ever more loose. I do not believe this nor wish it ...

...the farther a mathematical theory is developed ... unexpected relations are disclosed between hitherto separate branches of the science ... its organic character is not lost but only manifests itself with clarity.

Acknowledgements I

Amir Averbuch	Tel-Aviv University
Raphy Coifman	Yale University
Jerome Friedman	Stanford University
Trevor Hastie	Stanford University
Iain Johnstone	Stanford University
Art Owen	Stanford University
David Scott	Rice University
Bob Stine	University of Pennsylvania
Robert Tibshirani	Stanford University
Walter Willinger	ATT Shannon Labs

Acknowledgements II

Gregory Piatetsky-Shapiro	XChange
Edward Bosch	U.S. Army Topographic Eng. Ctr.
John Elder	Datamininglab.com
Al Inselberg	Tel Aviv University
Tzyy-Ping Jung	Salk Institute
Terry Sejnowski	Salk Institute
Jean-Luc Starck	Centre Europeen d'Astronomie
Peter Weinberger	Renaissance Technologies